Survival analysis project 1

FermaLogis Inc. Employee Attrition Analysis

**Group 2**

* **Suvir Gupta**
* **Reena Malipatil**
* **Sriram Rajagopalan**
* **Parmod Rathee**
* **Kai-Hsun Wang**
* **Deeksha Ramesh**

Contents

[Executive Summary 1](#_Toc486362765)

[Data Preprocessing: 2](#_Toc486362766)

[Methodology: 3](#_Toc486362767)

[Descriptive Analytics 4](#_Toc486362768)

[Survival Analysis 8](#_Toc486362769)

[Survival analysis - Year at Company 8](#_Toc486362770)

[Survival analysis – Total Working Years 9](#_Toc486362771)

[Summary 10](#_Toc486362772)

[Variable Selection(Stepwise:logistic regression) 12](#_Toc486362773)

[Modelling : Parametric(Life Reg) 13](#_Toc486362774)

[Conclusion 17](#_Toc486362775)

[Recommendations: 18](#_Toc486362776)

[Appendix 18](#_Toc486362777)

# **Executive Summary**

Employee attrition is one of the major concerns in all companies. FermaLogis is a pharmaceutics company facing this problem and we are hired to find the reasons for this cause. As per domain knowledge, there could be various reasons such as working overtime, job satisfaction level, business travel, employee working department etc. that could lead to employee attrition. FermaLogis dataset measures these variables and reports the proportions of employess that have left the company. In this survival analysis, we will profile the employees who are leaving the company and identify significant reasons for attrition.

We have identified that the employees leaving company are mostly the ones working overtime, and the people with job roles Laboratory Technicians, Sales Executive, Sales Representatives, and Research Scientists. We have also discovered that the attrition rate is high among low job levels and where the environment satisfaction level is low. Based on some of the information extracted from the initial analysis, survival models are built on year at the company and total working years. The survival plots show that the survival curve is steep from years 0 to 10, meaning that employees are more likely to leave the company at this period of their employment.

The content of this report describes the sampling, exploring, modifying and modeling process that was implemented to arrive at the conclusions reported above. The report emphasizes the results that are significant and have been achieved using standard statistical method.

# **Data Preprocessing:**

**Data Clean-up:**

On initial analysis of the dataset, we observed a few columns such as “X”, “Over 18”, “Employee Count”, “Employee Number” and “Standard Hours” that were either unique identifiers or not significant to build survival models. Hence, these columns were removed from the dataset.

**Variable Encoding:**

We observed the categorical variable “Business Travel” had different categories such as Non-Travel, Travel-Rarely and so on which gave information regarding the Employees and their travel frequency. We decided to encode this categorical variable into a continuous one by creating a new variable named Travel and assigning numerical values 0,1 and 2. Based on the different categories available, the following values were assigned to each group:

0 for Non-Travel

1 for Travel-Rarely

2 for rest of the categories which signifies that employees travel frequently

Similarly, we created another variable named Stock and assigned values “yes” and “No” based on the stock option level. For the all the values where Stock option level was greater than 1, the value “Yes” was assigned and “No” for the rest.

Finally, we created a variable “Censored” in order to take Censoring into account. If the Attrition is “Yes”, then the employee has left the company and hence the value “1” was assigned to the new variable “Censored”. If Attrition is “No”, then the exact time of employee leaving the company is unknown and hence the value “0” was assigned to the variable “Censored”.

**Aggregation:**

Since there were 40 columns which had details about bonuses given to the employees, we decided to aggregate them into a single variable named “Total Bonus” which gives the sum of the bonus received by each employee. To deal with the “NA” values, we decided to encode them to value 0 which means that the employee did not receive any bonus.

Final dataset after pre-processing had 76 variables 1470 observations in total.

# **Methodology:**

Once the data was pre-processed and ready for modeling, our initial approach was to determine the most significant variables that are contributing to the attrition of employees. In order to do this, we performed step-wise regression by backward elimination to determine the statistically significant variables which are affecting the attrition rate. For the Attrition event “Yes”, we performed step-wise regression and eliminated the insignificant variables in each step. The final results showed that there were 22 significant variables such as ‘Total Working Years’, ‘Years At Company’, ‘Years Since Last Promotion’ and so on that were contributing to Employee Attrition.

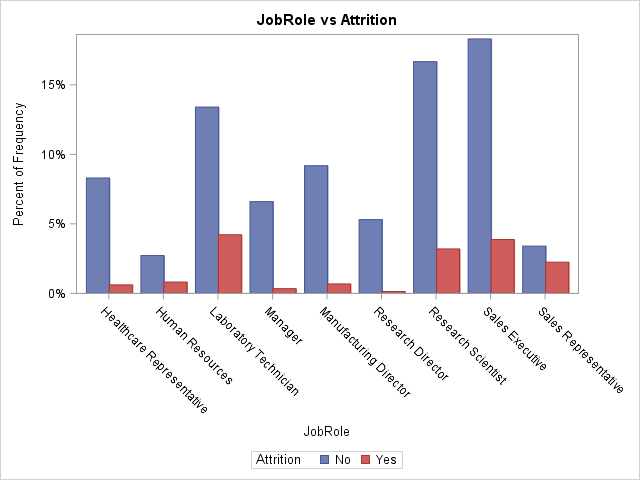
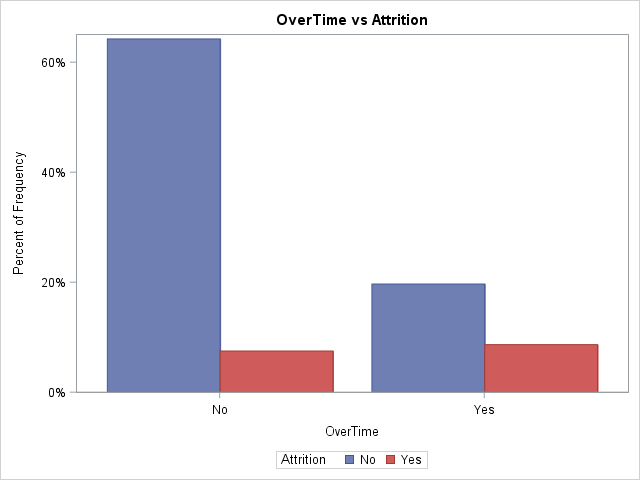
Since most of the significant variables were time variables, we decided to perform another step-wise regression on these to find out the top variables that are contributing to attrition. Once the most significant variables were determined, we explored these variables and observed the impact by means of visualization through plots. We used SGPLOT procedure in order to study the impact of each variable on Attrition rate through Bar charts, Frequency tables and other various plots available.

Additionally, to observe the impact of time variables, we used the LIFETEST method where events were grouped into intervals and the corresponding survival and Hazard plots were plotted. Finally, we built Accelerated Failure Time (AFT) model to study importance of co-variates though weighted average approach and determined the impact of each variable on survival time thereby analyzing the weights assigned to co-variates. To take into alternative distributions into account, we built several models with different distributions such as log-normal, Weibull etc. and compared the models to find the best fitting one through Log-Likelihood Estimates.

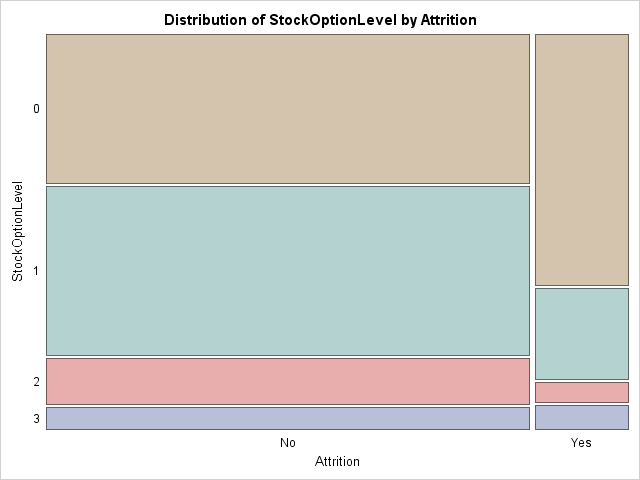
# **Descriptive Analytics**

According to the graph below, we can know some key factors about the tenures of employees. First, the graph “Over Time vs. Attrition” shows the employees who do overtime have higher attrition

when compared to other employees. Secondly, from the graph “Job Role vs. Attrition”. We can know that the highest attrition is in the Laboratory Technicians, Sales Executive, Sales Representatives, and Research Scientists. Research Directors and Managers have the lowest attrition rates.

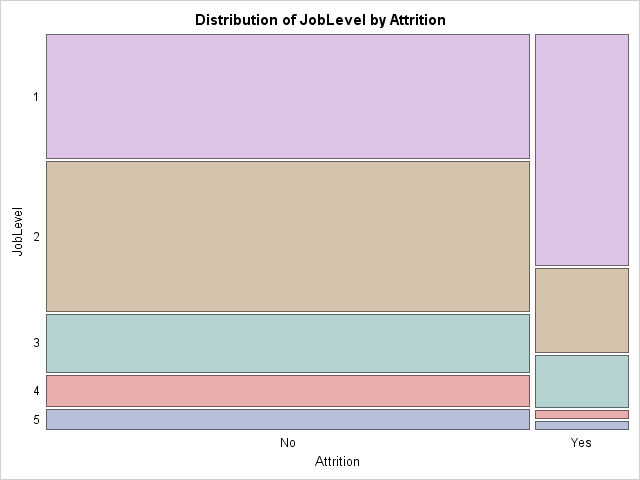


To understand more about the data insight, we use the mosaic plot to observe other factors influencing the turnover of an employee. From the mosaic plot (a), we know that the lower the stock option level, the higher the attrition rate. Therefore, giving stock options is a good way to motivate employees to stay in company.



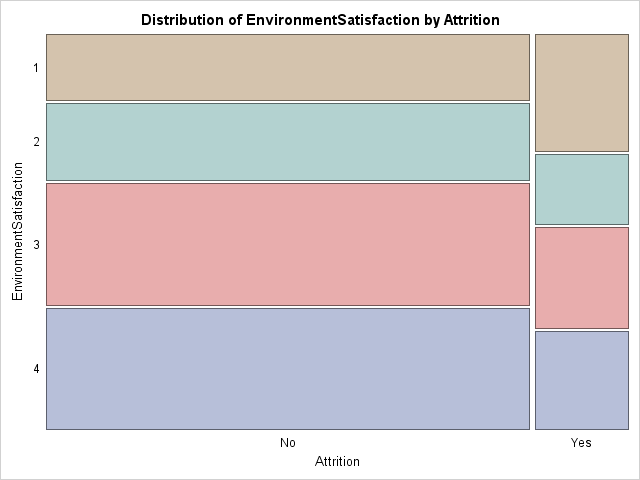
Mosaic(a)

From the mosaic plot (b), we know that the lower the job level, the higher the attrition rate. The reason junior employees have higher leaving rate could be attribute to the low opportunity cost of job change and the organizational culture conflict.



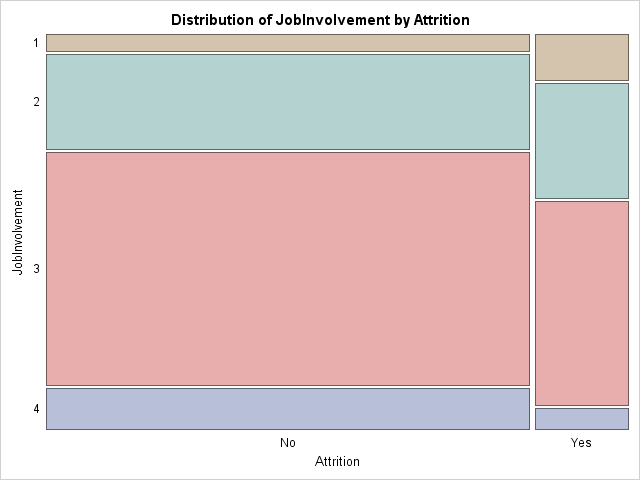
Mosaic(b)

From the mosaic plot (c), we know that the lowest the environment satisfaction, the highest the attrition rate. However, in the environment satisfaction level 2, 3, and 4, there is no significant correlation between environment satisfaction and attrition rate.



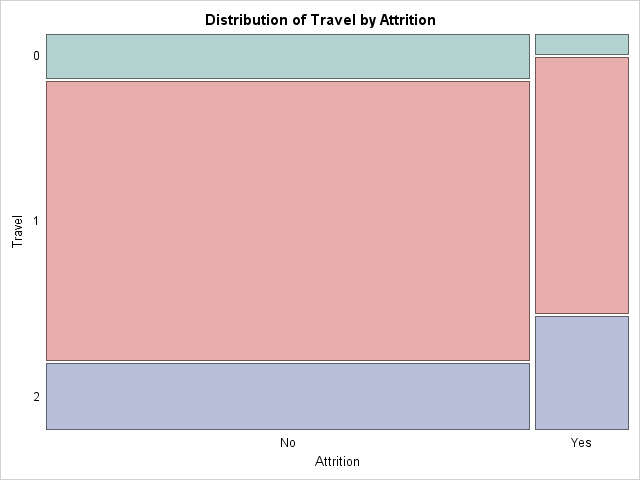
Mosaic(c)

From the mosaic plot (d), we know that the job involvement level 2, and level 3 have the highest the attrition rate. The job involvement level 4 has the lowest the attrition rate.



Mosaic(d)

From the mosaic plot (e), we know that the lower the frequency employee travel, the lower the attrition rate.



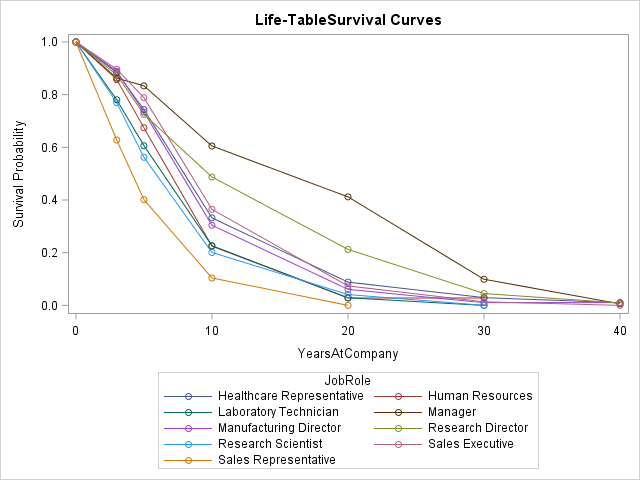
Mosaic(e)

After understanding the variables of data, we implemented survival analysis method to understand the underlying factors influencing the turnover of an employee. Also, we developed the models to predict the possible outcome of the tenure of an employee.

# **Survival Analysis**

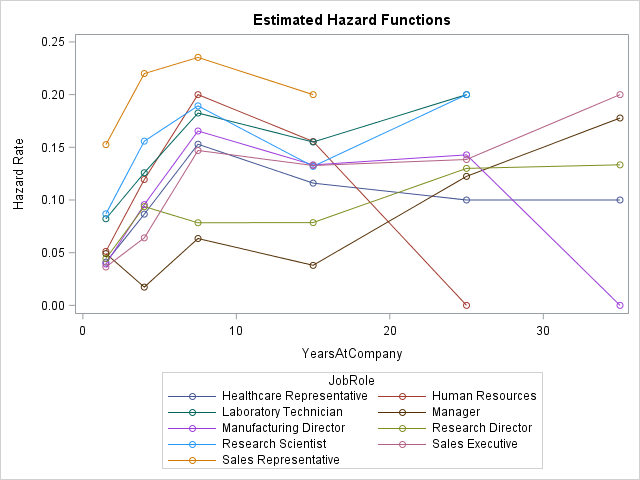
Based on the descriptive analysis above, we tried to focus more on employee background and job profile. Therefore, we developed a survival analysis model for the employee data. In order to better understand the dynamics of employee retention, we developed different models by comparing the job role with working years and the time employee at company.

## **Survival analysis - Year at Company**



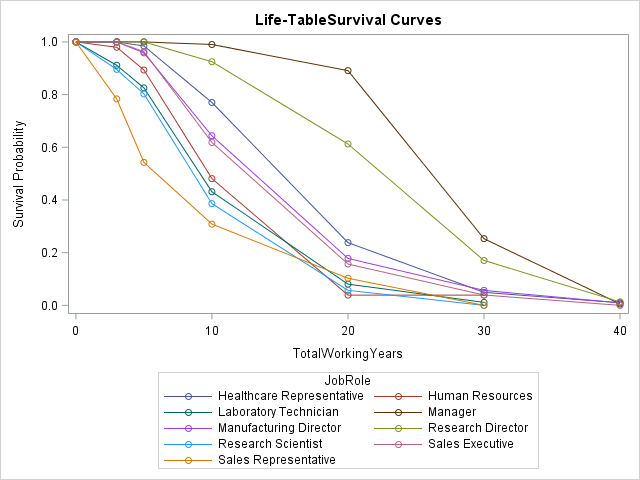
The survival probability plot above shows the survival probability based on the years at company.

As we can see, the sales representative has the lowest survival probability. On the contrary, manager and research director have the highest survival probability. Moreover, we can know that the year at company between 0 to 10 is steep. It means that the employees are more likely to leave the company in the period between year 0 to year 10. After year 10, the leaving rate will be more stable.

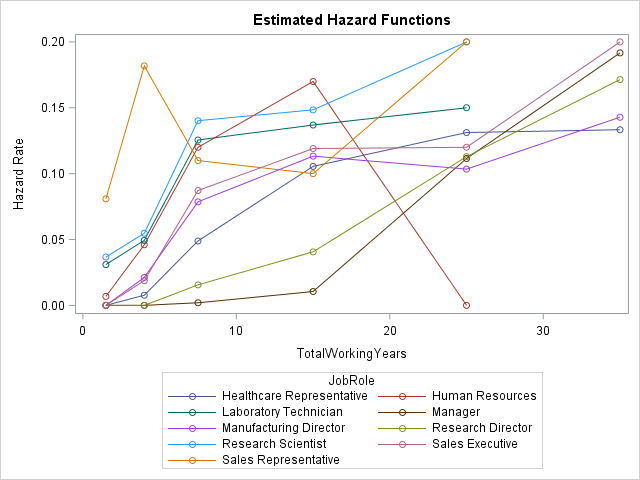


The hazard for Sales Representative is the highest from the first few years. During year 6 to year 15, the hazard rate for all positions decline. The hazard rate for manager and sales executive increases sharply after year 15. However, the hazard rate for human resource and manufacturing director decreases sharply after year 15 and 25 respectively.

## **Survival analysis – Total Working Years**



The survival probability plot above shows the survival probability based on the total working years. The survival probability plot above shows the sales representative still has the lowest survival probability. However, after year 15, human resource and research scientist have the lowest survival probability. In this model, manager and research director still have the highest survival probability.



In this model, the hazard rate for all positions are increasing, excluding sales representative and human resource. The hazard for Sales Representative is high from the first three years, but it decreases sharply from year 3 to year 15. The hazard rate for human resource increases sharply from year 0 to year 15. However, the hazard rate for human resource drops sharply after year 15.

# **Summary**

Here is the summary of our finding:

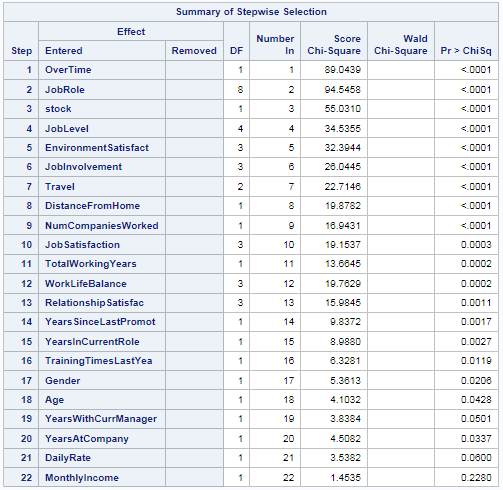
|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Who are leaving? | Possible reasons | Recommendation |
| Over time | Employees who work overtime have a higher risk of leaving the company than others. | Workload is too high for the employees and the employees are not satisfied with the overtime pay. | Appropriate breakdown of the work and improved overtime pay. |
| Job Role | The highest attrition is in the Laboratory Technicians, Sales Executive, Sales Representatives, and Research Scientists. | The sales department and the research department are the most important departments in healthcare industry. Thus, the employees in these two departments are more likely working under pressure. | Improve job related training accessibility. It will engage employees and they can set their smart goals for every year. |
| Stock Options | The lower the stock option level, the higher the attrition rate. |  | Giving stock options is a good way to motivate employees to stay in company. |
| Job Level | The lower the job level, the higher the attrition rate. | Junior employees have higher leaving rate could be attribute to the low opportunity cost of job change and the organizational culture conflict. | Employee motivation and job training for higher roles should be implemented. |
| Work Travel | The higher the frequency employee travel, the higher the attrition rate. | Work travel is extra workload for employees. | Strategic hiring process and resourcing. |

When is the biggest danger for employees to leave?

|  |  |  |
| --- | --- | --- |
| **Job Role** | **When (Years in Company)** | **When (Working Year)** |
| Sales Representative | 0~3 | 0~3 |
| Sales Executive | 3~6 | 3~6 |
| Human Resource | 0~6 | 3~6 |
| Manager | 15~25 | 15~25 |
| Research Director | 0~3 | 15~25 |
| Healthcare Representative | 0~6 | 3~6 |
| Laboratory Technician | 0~6 | 3~6 |
| Manufacturing Director | 0~6 | 3~6 |
| Research Scientist | 0~3 | 3~6 |

# **Variable Selection(Stepwise:logistic regression)**

To select the significant covariates for survival analysis we run stepwise logistic regression taking ‘*Attrition’* as the response variables



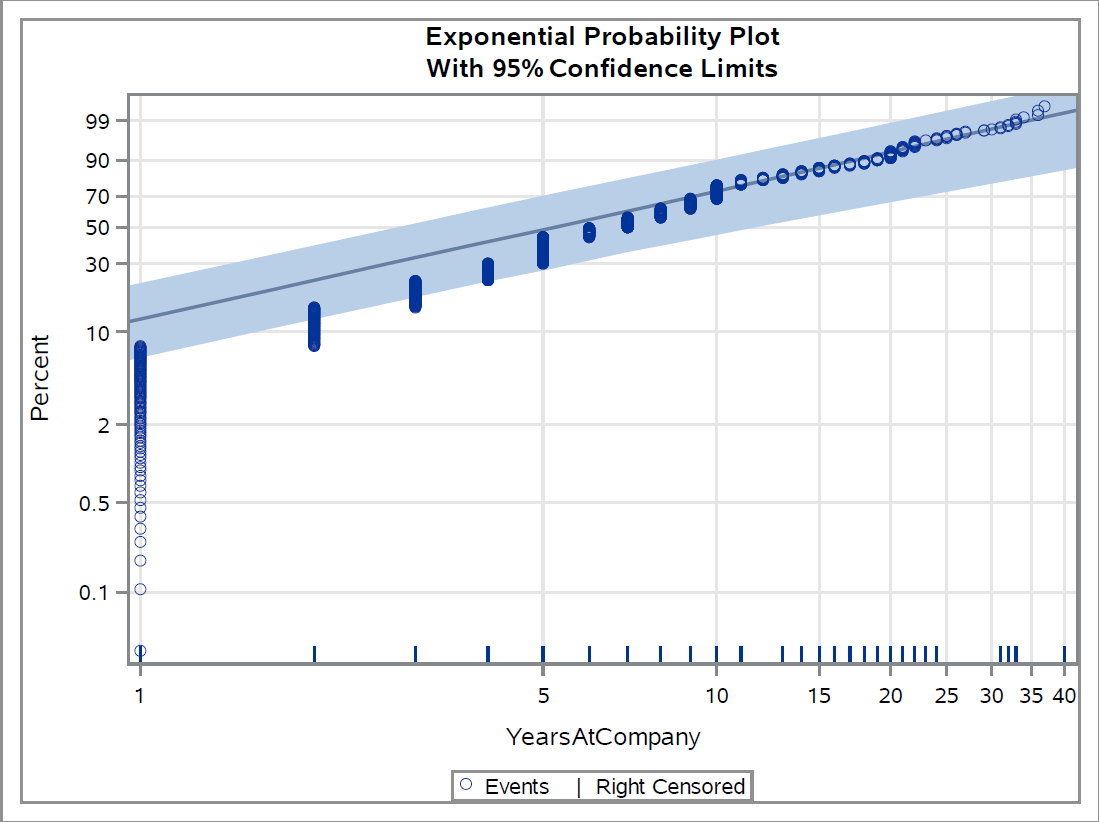
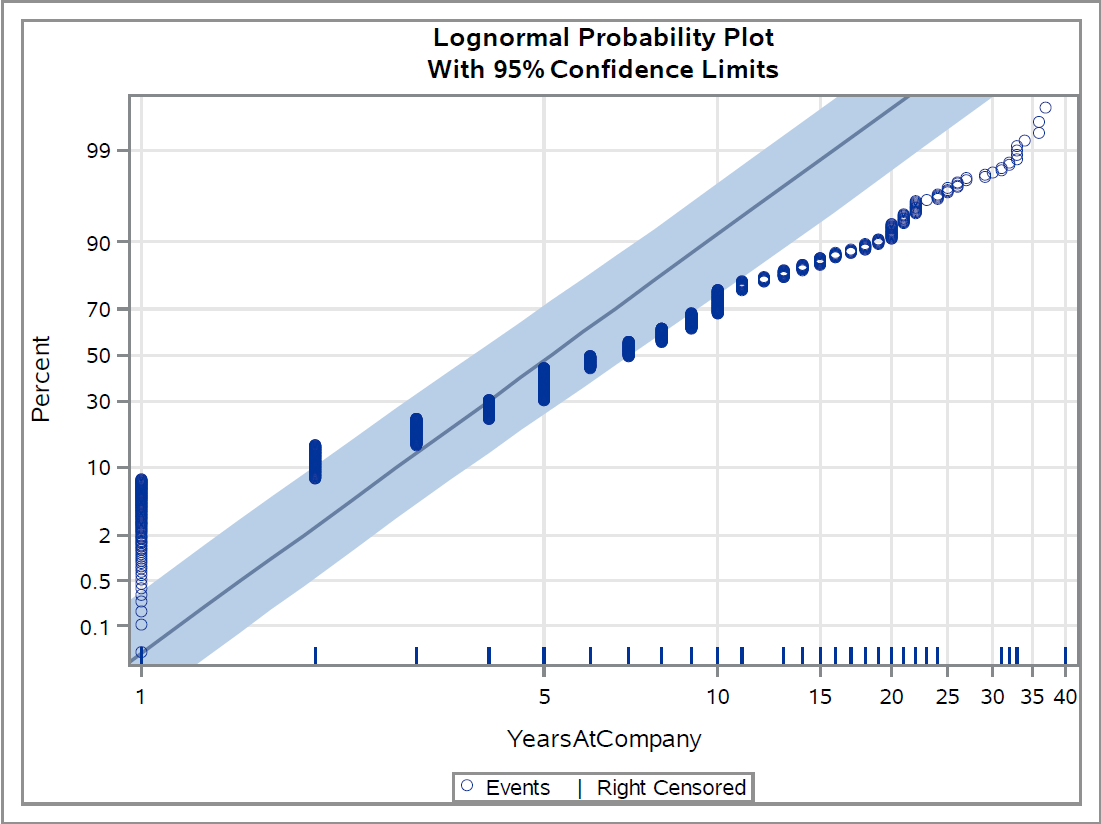
Above are the covariates that can be used for the survival model. We may further remove them from the survival model selection based on low p-value and can retain some based on business value like YearsWithCurrManager, YearInCurrentRole.

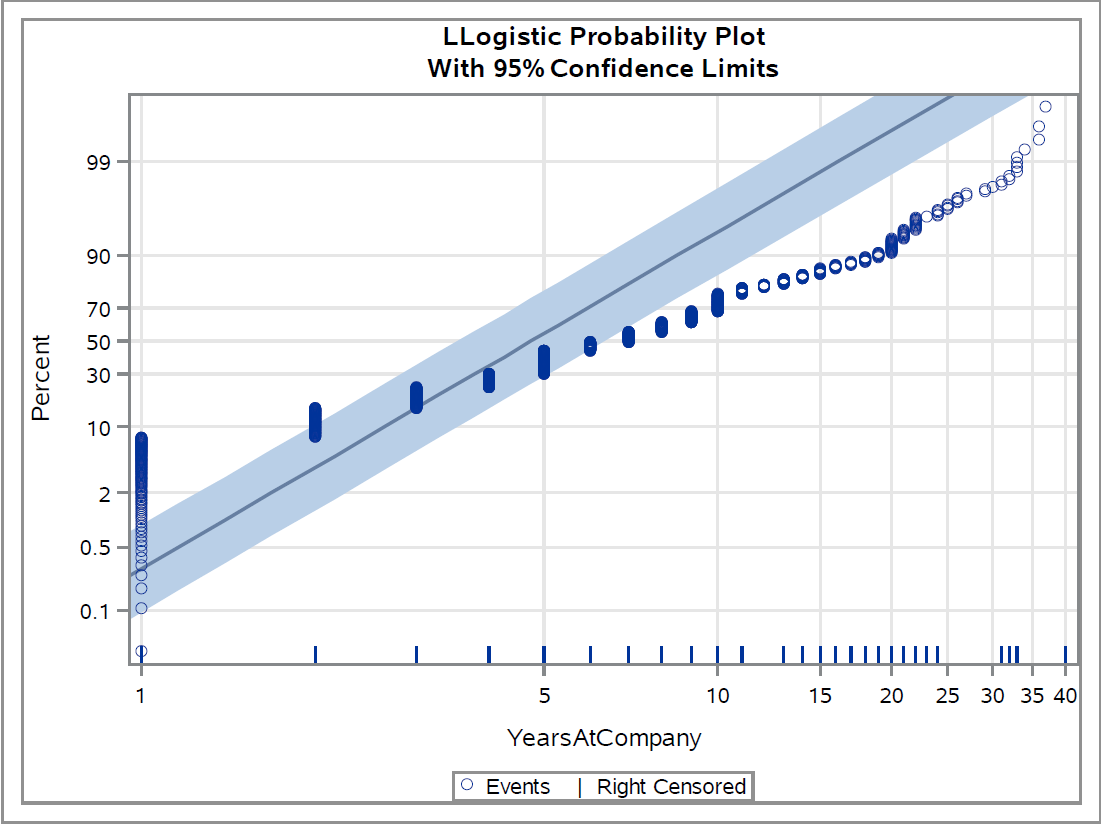
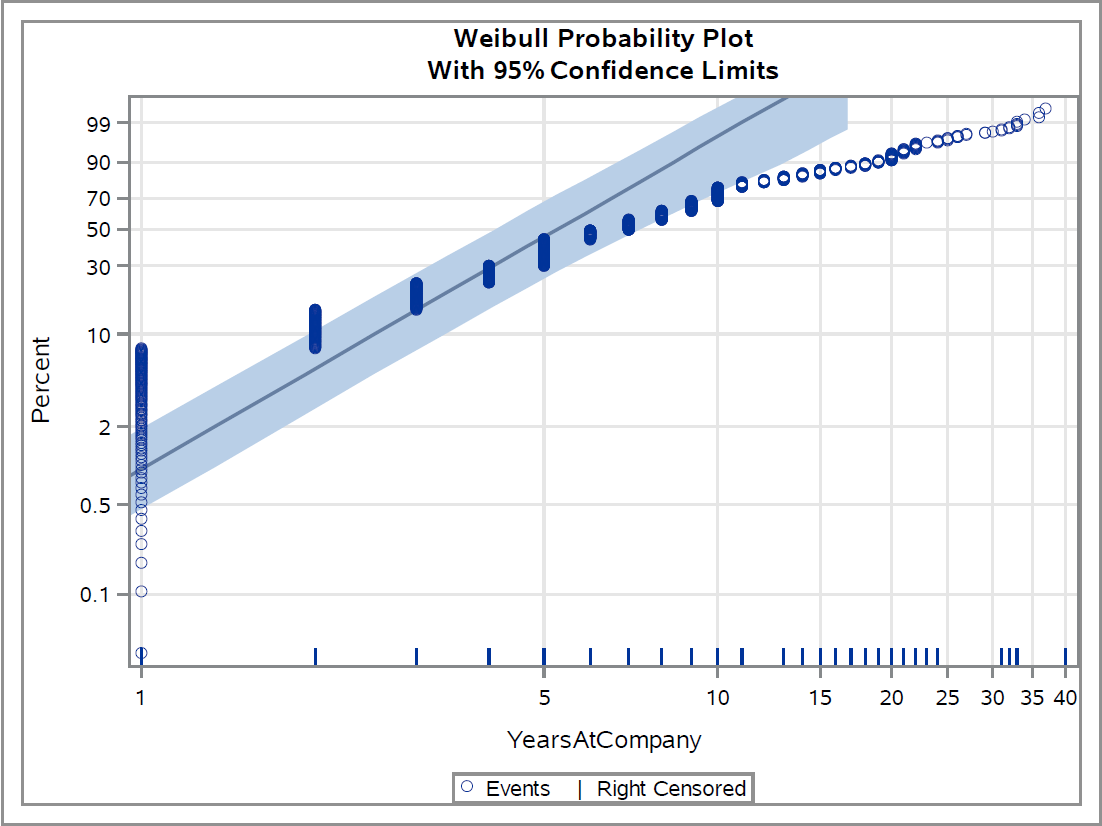
# **Modelling : Parametric(Life Reg)**

We are taking top 15 covariates selected based on the stepwize selection criteria of the logistic regression using attrition as the response variable.

For survival analysis *Attrition* is the time event for which censored = 0 as explained in data preprocessing. We built parametic model on multiple distributions like Lognormal, Exponential, Weibull, Llogistic and gamma distribution.

Below are the probability plots for all distributions. Gamma distribution is excluded as it did not converge.



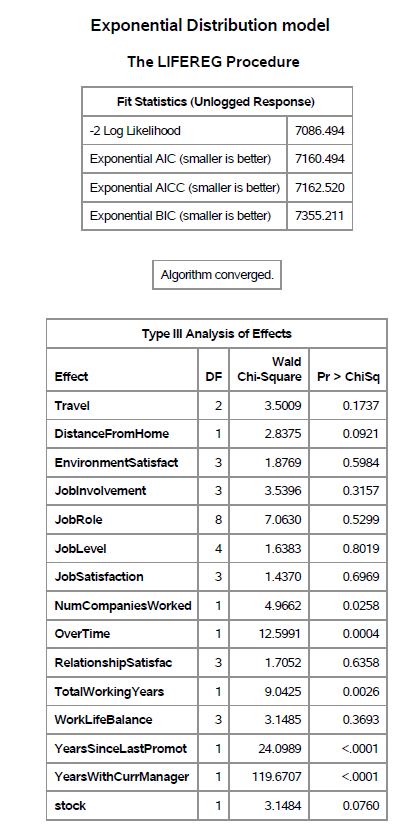
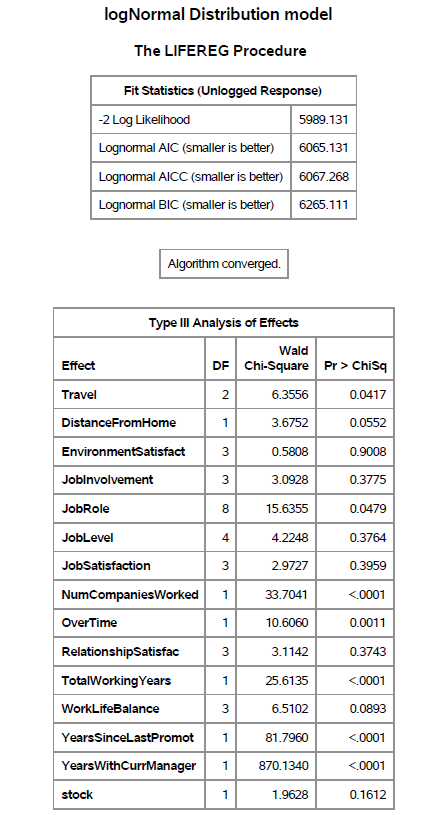


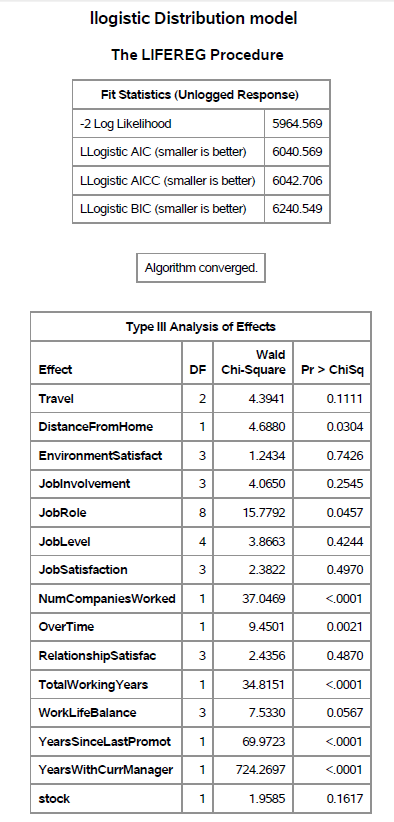
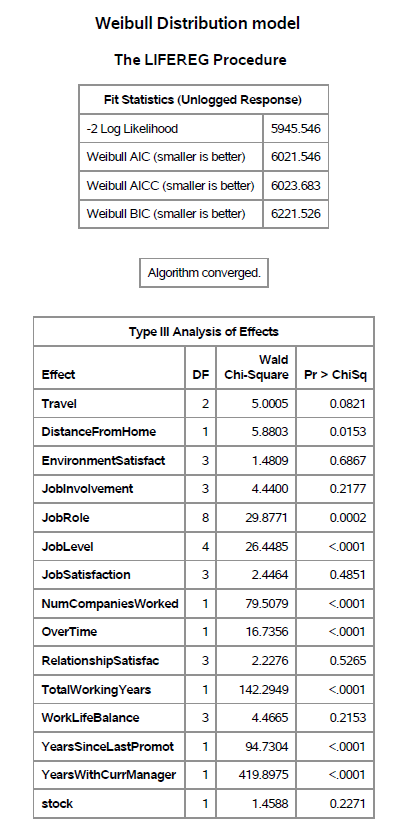
Having a look at the graphs it is visible that exponential and Weibull seems to predict the best among all the distribution.

Comparing Log likelihood of all models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | LogNormal | Exponential | Weibull | Loglogistic |
| Log likelihood | -941.3043243 | -1489.985557 | -919.5118298 | -929.0229488 |

**Fit statistics off all distributions**





**Comparison with null model**

Comparing the log likelihood value with the null model for all distribution all p-values are 0 implies that all models are better fit compared their null counterpart.

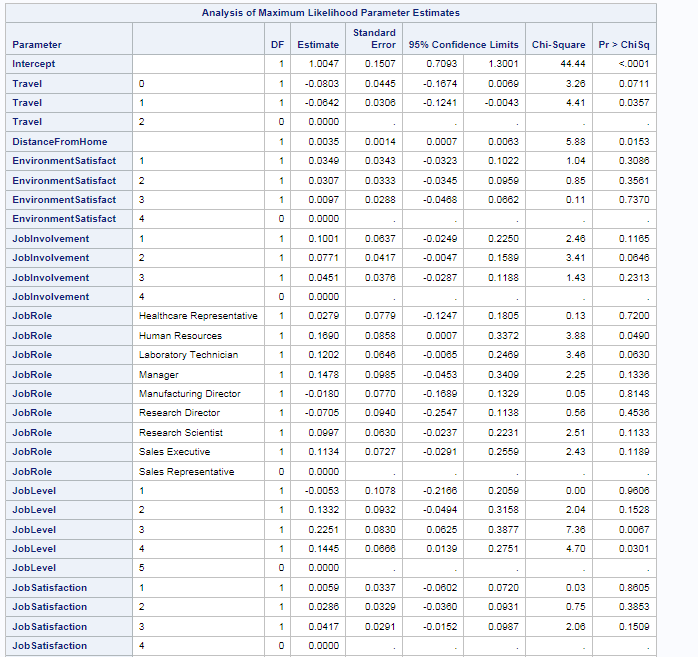
**Best Model**

Hence model selection is done based on the visual analysis of the probability plot. Even though the log likelihood values of Exponential distribution is low, but probability plot shows its better prediction model for YearsAtCompany>2. Overall model prediction is best for Weibull Model.

It can be inferred from the plots that we can use Exponential model to predict survival time for employees with years in company greater than two and for the rest of the prediction we can use Weibull model.

There are two best model as explained above but we consider Weibull with best overall statistics to be considered for parameter significance.

**Parameter significance of Weibull model**





Percentage contribution of each parameter to the survival time can be calculated using the estimate column in the table above using below formula.

* Coefficient is positive; if coefficient increases, expected survival time increases.
* Coefficient is negative; if coefficient increases, expected survival time decreases.
* For each unit of parameter increase results in % increase(or decrease) in the expected survival time.

Hence Estimate column represents the factor’s contribution positively or negatively to survival time.

# **Conclusion**

Based on our analysis the employees leaving the company are the ones not owning stocks, not satisfied with the job involvement, and not having work life balance. It is also seen that the employees with less experience are leaving the company for a higher compensation in other companies. Employees with higher experience are leaving because to get rid of overtime as shown in the survival model results. Employee attrition is highest in the second and third year of employment. This shows that young employees tend to move out more. There is a difference in attrition between different employee groups. Employees who are involved in operations highly are less satisfied with their job due to which their attrition rate is higher.

# **Recommendations:**

* Improve job related training availability and encourage employees which are on job roles like Laboratory Technicians, Sales Executive, Sales Representatives. It will engage them in the organization and break their monotonous work routine with a new goal set in mind.
* To avoid overtime, mangers must allocate the work to resources optimally.
* Strategic hiring process and resource allocation to reduce frequent travel.
* Provide sufficient travel and overtime allowance to the employees.
* Provide stock options to the employees for better engagement in the FarmaLogis.

# **Appendix**

Please find the below SAS file which contains code to obtain the above explained results.

